

# HPSS Connections to ESGF: BASEJumper

Presented to ESGF Face to Face 2016

Sam Fries  
Analytics and Informatics Management Systems  
LLNL

December 8, 2016



# Model Output is big. Really big.

---

- CMIP6 is estimated to clock in at 60PB
- 5 years of low resolution ACME model output is ~1TB
- Only getting bigger.
- Spinning disk capacity is not growing fast enough.

# Tape to the rescue!

- Tape archives deployed at most computing facilities

Solid State	Hard Disk	Tape (LTO6)
\$250 / TB	\$58 / TB	\$12.50 / TB

Note: Price of single drive on Newegg; YMMV

- DOE compute facilities have “High Performance Storage Systems”

# Tape is hard ☹️

---

- HPSS goes down
- Transfers fail
- Discoverability is nil

# ESGF + Tape?

---

- How do we publish the data?
- How do we get the data?
- How do we not DDOS compute facility resources?

# BASE: Berkeley Archival Storage Encapsulation

---

- Library for automating use of HSI/HTAR
- C++ with Python API wrapper
- Made by Alex Sim, from the SDM group at Lawrence Berkeley

# Publishing the Data

---

- Use BASE to extract necessary file metadata
  - Filesize
  - Checksum

# ESGF + Tape?



How do we publish the data?

- How do we get the data?
- How do we not DDOS HPSS systems?



# Getting the Data

---

- Use BASE to retrieve files
- Provide cache of transferred files with some minimum duration guarantees

# ESGF + Tape?



How do we publish the data?

How do we get the data?

- How do we not DDOS HPSS systems?

# ESGF + Tape?



How do we publish the data?

How do we

- How do we



# THE CALL IS COMING FROM INSIDE THE HOUSE

---

- Use two separate pieces
  - Web frontend for dealing with ESGF
  - Daemon backend for dealing with HPSS
- Daemon lives inside compute facility firewall
- Asks frontend for what files to transfer

# ESGF + Tape?



How do we publish the data?

How do we get the data?

— How do we get data past firewalls?

- How do we not DDOS HPSS systems?

# ESGF + Tape?



How do we publish the data?



How do we get the data?

How do we get data past firewalls?

- How do we not DDOS HPSS systems?

# ESGF + Tape?

---



How do we publish the data?



How do we get the data?



How do we get data past firewalls?



How do we not DDOS HPSS systems?

# BASEJumper Overview

---

- Public-facing web API (frontend)
  - Authenticated through OpenID, access control via SAML
- Daemon for interacting with HPSS via BASE
  - Tasks scheduled by frontend API
- Scripts that interact with BASE and the frontend to retrieve metadata for the publisher



# Frontend In Depth


- Files show up in search
- “basej” link queues transfer
- Transfer appears in user profile with status
- Status is updated as transfer happens
- Email sent at completion with download link

1. **hpss-demo.test.v1\_1.atmos**  
Data Node: pcmdi11.llnl.gov  
Version: 4  
Total Number of Files (for all variables): 1  
[\[ Show Metadata \]](#) [\[ Hide Files \]](#) [\[ THREDDS Catalog \]](#) [\[ WGET Script \]](#) [\[ VISUS IDX \]](#)

Total Number of Files: 1

**ed3bef5bf3cc46584943b45ae4cdd45e3e319d7558529796e1f35a3d7d683424**

1 Checksum: 96b3b5fa2ec6f38fb8f37e8d18ef937558230537095c6015882f02bece4a31ea [basej](#)  
Size: 1718908  
[\[ More File Metadata \]](#)

 [Add to Data Cart](#)

HPSS Transfers	
/home/a/asim/asim.hash.txt	In Queue

HPSS Transfers	
/home/a/asim/asim.hash.txt	55% Transferred

# Daemon In Depth

---

- Can be deployed behind computing facility firewall
- Calls the frontend's API to determine what work to do
- Error-tolerant transfer process
- Customizable data transfer mechanism from daemon to frontend

# Status & Future Work

- Deployed on a development node currently
- Few rough edges to clean up
  - Automate publishing process
    - Publish & Archive Script:
      - Extract search facets
      - Calculate checksum
      - Archive in HPSS
      - Setup file permissions
      - Publish through Ingestion API and BASEJumper Frontend
  - Migrate authentication/access control to oauth
  - Usability improvements and features to add:
    - Multiple files in one transfer
    - Accessing files from within HTAR archive
    - Providing download statistics for ESGF Dashboard

